ÉCOLE NATIONALE DES CHARTES

Paul, Hector Kervegan

Traitement, exploitation et analyse d'un corpus semi-structuré : le cas des catalogues de vente de manuscrits

Mémoire pour le diplôme de master

Technologies numériques appliquées à l'histoire

2022

Résumé

Introduction

J'ai choisi de structurer mon mémoire autour de plusieurs questions connexes, qui, à différents degrés, se retrouvent tout au long du développement :

En quoi la nature semi-structurée du corpus permet d'en automatiser le traitement ? Comment produire des informations normalisées et exploitables à partir d'un corpus textuel semi-structuré ? En quoi ce traitement et la traduction des documents vers d'autres formats et d'autres médias impacte leur réception ? Quels sont les choix techniques qui influencent cette réception ?

Les deux premières questions, d'orientation plutôt technique, forment la colonne vertébrale pour le mémoire ; elles lient deux aspects centraux : la nature du corpus et la manière dont sa structure permet toute la chaîne de traitement. Par « semi-structuré », j'entends que, à un niveau distant, toutes les entrées de catalogue suivent la même structure ; des séparateurs distinguent les différentes parties, et les informations sont souvent structurées de manière semblable pour chaque manuscrit vendu. Cela permet un traitement de « basse technologie » (*low-tech*) en évitant d'entraîner de lourds modèles de traitement du langage naturel (ce qui aboutirait à des solutions complexes, difficiles à maintenir et à faire évoluer et relativement opaques dans leur fonctionnement). À l'inverse, un corpus semi-structuré peut être traité en déduisant une « structure abstraite », que chaque entrée de catalogue partage. Il est alors possible de mettre en place des solutions techniques plus faciles, pour un résultat de qualité équivalente. Produire des « informations normalisées et exploitables » implique de traiter le corpus en cherchant des réponses à des questions de recherche précises – dans le cadre de mon stage, une question centrale a été de chercher à isoler les facteurs déterminant le prix d'un manuscrit.

Les deux dernières questions, au premier abord plus théoriques, me semblent centrales, notamment à la troisième partie de ce mémoire. Numérisation, traitement informatisé et diffusion sur le web ne sont pas des opérations neutres, mais un ensemble de « traductions » des documents originels. Ces processus comportent une part de choix conscients, qu'il s'agit de mettre en avant. Par exemple, on considère que la majorité des documents vendus ont pour titre l'auteur.ice du document. Cette personne n'est cependant pas toujours mentionnée, et des documents peuvent être nommés d'après un lieu, un évènement ou un thème (la révolution française, par exemple). Ces « traductions » des catalogues sont relativement discrètes tout au long de la chaîne de traitement (où le

format dominant est la `TEI`, qui garde une relation d'équivalence avec le texte). C'est lors du passage au site web que ce processus de traduction devient plus évident, et, potentiellement, plus problématique. On y abandonne la référence au document originel (les catalogues numérisés ne sont pas accessibles en ligne par un.e utilisateur.ice), le catalogue n'est plus la manière privilégiée d'accéder aux items vendus... De plus, la construction d'un site web implique la conception d'une interface et, dans notre cas, la production d'une série de visualisations intégrées au site. Le passage au site web remet aussi en cause la hiérarchie habituelle entre ingénierie et recherche : la conception d'un site ne répond pas à une question scientifique, mais elle soulève ses propres questions. Loin d'être anodines, ces problématiques de design déterminent la construction et la réception des savoirs. Il est donc important, je pense, de problématiser ces questions de visualisation et de design.

Première partie

Du document numérisé au XML-TEI : nature du corpus, structure des documents et méthode de production des données

Chapitre 1

Présentation du corpus

Ce chapitre est dédié à une présentation des documents traités dans le cadre du projet MSS : nature, quantité de documents (et d'entrées individuelles), dates et différents types de catalogues vendus. On pourra également représenter la répartition des ventes par an grâce aux graphiques produits pour le site web avec Plotly (à moins que cela ait plus de sens en troisième partie). Ce chapitre s'appuie sur les mémoires effectués par d'ancien.ne.s stagiaires de Katabase, qui ont déjà beaucoup analysé la nature et les enjeux du corpus [1].

1. Lucie Rondeau du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Charavay*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : `https://github.com/lairaines/M2TNAH` (visité le 13/06/2022) ; Caroline Corbières, *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH` (visité le 13/06/2022) ; Juliette Janès, *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

Chapitre 2

Production des données : océrisation et structure des documents traités

Cette partie s'attache autant à présenter le processus d'océrisation (qui est déjà bien établi et ne constitue pas le cœur de mon stage) que la structure des documents. Alors que le chapitre précédent s'intéresse aux catalogues dans leur ensemble, ici, on étudie le corpus au niveau de la page et de l'entrée individuelle. En effet, l'océrisation repose sur la segmentation, et donc sur l'établissement d'une structure « abstraite » d'une page (c'est-à-dire, d'un découpage de la page en zones).

Chapitre 3

Du texte à la TEI : méthode de transformation des documents océrisés en fichiers XML-TEI valides

Après une étape d'océrisation via eScriptorium, le texte extrait des PDF peut être exporté soit en texte brut, soit en XML Page ou Alto. Ces formats s'attachent à garder une relation entre le XML et le document numérisé (les zones de texte sont indiquées, chaque ligne est dans une balise...). Cependant, l'unité intellectuelle centrale à la suite du projet, ce n'est pas la page numérisée, mais l'entrée de catalogue. Un format plus complexe que le XML d'eScriptorium est donc nécessaire. Assez logiquement, la suite du projet s'appuie sur une traduction des catalogues en TEI.

Jusqu'à maintenant, cette transformation était faite par GROBID-Dictionnaries et des feuilles XSL. Le projet GROBID-Dictionnaries étant maintenu de façon assez opaque, le choix a été fait de remplacer cet outil par une solution plus simple (qui ne passe pas par le *machine learning*) et modulable. En s'appuyant sur la nature « semi-structurée » du corpus, il est possible de séparer les différentes parties du texte en identifiant des séparateurs. Ce chapitre s'intéresse donc à la transformation de Alto vers la TEIà l'aide d'*expressions régulières*. En changeant d'unité intellectuelle – ou en choisissant de privilégier une unité au sein des catalogues plutôt qu'une autre (page, série d'entrées...) –, on prend de la distance et on interprète le catalogue papier.

Deuxième partie

Normalisation, enrichissements et extraction d'informations : une chaîne de traitement pour des données semi-structurées

Chapitre 4

Homogénéiser et normaliser un corpus complexe

Ce chapitre s'intéresse à la manière dont les fichiers `TEI` sont traités afin de pouvoir ensuite en extraire des informations. C'est directement grâce la structure des entrées (et grâce à la nature « semi-structurée » des catalogues) qu'est possible le traitement automatisé des documents. Cette partie s'attache également à rappeler les questions de recherche qui sous-tendent la normalisation des documents (ajouter plus de structure au document `TEI` pour l'exploiter plus facilement, uniformiser la notation des tailles et des dimensions des documents...). Cette partie sera probablement relativement brève, puisque l'étape de normalisation des données a déjà été faite par d'autres stagiaires et documentée d'autres mémoires.

Chapitre 5

Extraction d'informations

Ici sont décrits le processus et les objectifs de l'extraction d'informations à partir des fichiers `TEI`. Des données sont extraites pour chaque entrée de catalogue (un travail largement effectué par A. Bartz, que j'ai légèrement mis à jour) : prix (dans la monnaie de l'époque et en francs constants), date de vente normalisée, nom de l'auteur.ice et description du manuscrit... Un second processus d'extraction produit des données pour chaque catalogue de vente : titre du catalogue, date de vente, nombre d'items vendus, prix minimum, inférieur et maximum, prix moyen et médian, variance... En même temps que ces processus d'extraction d'informations, un script de conversion des monnaies (françaises et étrangères) en francs constants 1900 a été élaboré. En annulant l'effet de l'inflation, les francs constant permettent d'étudier l'évolution réelle des prix.

L'extraction d'informations pour les entrées individuelles permet surtout de faire une réconciliation des manuscrits vendus (c'est à dire, de retrouver les items vendus plusieurs fois). Le deuxième processus permet de produire des données statistiques sur l'évolution du cours et du volume du marché des manuscrits (nombre d'items en vente, évolution des prix) ; c'est à partir de ces informations normalisées que sont construites les visualisations intégrées au site de Katabase.

Chapitre 6

Vers une étude des facteurs déterminant le prix des documents : alignement des entrées du catalogue avec Wikidata et exploitation de données normalisées

Ce chapitre est construit autour d'une question de recherche : comment produire des informations exploitables pour une étude économétrique à partir d'un corpus textuel semi-structuré ? Un des objectifs du projet est de faire l'étude des facteurs déterminant le prix d'un manuscrit. Pour faire cette étude, il faut obtenir, pour chaque entrée du catalogue, un certain nombre d'informations normalisées. Le travail d'extraction de données présentes dans les catalogues a déjà été fait par de précédent.e.s stagiaires. Ces données sont principalement quantitatives : prix des manuscrits, dimensions et nombre de pages, date de création. Il est nécessaire de compléter les informations par des données qualitatives et d'enrichir les données disponibles avec des sources extérieures. Pour ce faire, il a été choisi d'aligner le nom des auteur.ice.s des manuscrits avec des identifiants Wikidata ; dès lors que l'on a un identifiant Wikidata, il est possible de récupérer automatiquement des informations sur les personnes via `SPARQL`. Le choix de travailler uniquement sur les noms, et non sur la description des documents, a deux motivations :

— Les noms de personnes (et la manière dont elles sont décrites) constituent la partie la plus normalisée des documents. La description des manuscrits est plutôt en « texte libre ». Dans la continuité avec le reste du projet, nous sommes resté dans une approche « basse technologie », qui consiste à s'appuyer majoritairement sur des solutions techniquement simples. C'est pourquoi nous avons préférer traiter

les noms avec des tables de correspondance [1] et des *expressions régulières*, plutôt que de faire du TAL sur la description des documents.

— Toutes les informations « simples » (données quantitatives facilement normalisables : dates etc.) ont déjà été extraites des descriptions des manuscrits.

Ce travail d'enrichissement a été fait en deux temps.

La première étape, et la plus difficile, est l'alignement avec Wikidata. Cela demande d'extraire un ensemble d'informations à partir du nom de la personne et de la description de celle-ci (nom, prénom, titre de noblesse, occupation, dates de vie et de mort...). À partir de ces informations, stockées dans un dictionnaire, un algorithme construit successivement différentes chaînes de caractères à rechercher sur l'API de Wikidata. L'objectif est que le premier résultat recherché sur Wikidata soit correct. Sur un jeu de test, le score F1 [2] obtenu est de 68%. Un relecture « manuelle » des résultats est donc nécessaire.

La deuxième étape, nettement plus simple, consiste à lancer des requêtes Wikidata sur les identifiants récupérés afin de récupérer des informations sur les auteur.ice.s des manuscrits (cette partie du travail est encore en cours) pour enrichir nos données.

1. C'est à dire, des tables qui permettent de normaliser la manière dont les informations figurent dans les catalogues, et donc de remplacer des termes « vernaculaires » par leurs équivalents utilisés par Wikidata

2. Le score F1, ou *F-score*, est la moyenne harmonique de la précision (vrais positifs par rapport au total de résultats obtenus) et du rappel (nombre de résultats positifs par rapport au total de résultats positifs). (*Précision et rappel*, Wikipedia. L'encyclopédie libre, 2022, URL : `https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel` [visité le 13/06/2022])

Troisième partie

Après la TEI : l'application web *Katabase*, interface de diffusion des données

Chapitre 7

Nouvelles présentations du corpus : l'interface web Katabase

Ici, on s'intéresse à la manière dont le site web de Katabase permet de proposer une visualisation différente du corpus :

- création de différentes manières de naviguer dans le corpus et mise en lien entre le corpus et le projet Katabase
- création d'un algorithme de recherche et de réconciliation des manuscrits, qui recoupe les différentes ventes afin d'identifier si un manuscrit a été vendu plusieurs fois
- visualisations de données et développement de graphiques interactifs sur la page d'index des catalogues et pour chaque catalogue

Les deux premières parties ont été réalisées par Alexandre Bartz. Elles ne concernent pas le cœur du projet. J'y reviens cependant pour mettre en avant que la création d'une interface web implique l'éloignement du document originel. Le site internet marque avant tout un éloignement intellectuel avec les documents : le catalogue n'y est plus l'unité intellectuelle dominante, alors qu'il restait l'un des critères structurants des fichiers `TEI` (un fichier représentant un catalogue). Sur le site web, on peut accéder directement aux éléments vendus, sans avoir à passer par les catalogues. L'éloignement du document originel est aussi technique, puisqu'on abandonne totalement la `TEI` et les formats sémantiques au profit de formats de présentation (`HTML`), tout en s'appuyant largement sur des formats « techniques » et pratiques (le `JSON`, qui est simplement un moyen de rendre des données rapidement accessibles).

Dans cette partie, je m'intéresserai plus spécifiquement aux problématiques techniques et scientifiques liées à la visualisation de données, qui ont constitué une des missions centrales de mon stage :

- Qu'est-ce que l'on cherche à montrer avec ces visualisations ?
- Comment intégrer ces visualisations de manière « élégante » et efficace au site ? Par exemple, 7 types de graphiques sont produits pour l'index des catalogues. Il

faut donc trouver un moyen de tous les présenter, sans pour autant alourdir excessivement la page. La visualisation doit également être pensée en concertation avec le reste du design du site web (largement mis à jour pendant mon stage), et s'intégrer à sa charte graphique.

— Comment construire ces visualisations ? Comment faciliter la lecture d'informations complexes ? Quel impact les visualisations ont-elles sur la perception des informations ? Pour ces problématiques techniques, je pense (entres autres) m'appuyer sur le mémoire de Ségolène Albouy [1].

1. Ségolène Albouy, *Médiation des données de la recherche. Élaboration d'une plateforme en ligne pour une base de tables astronomiques anciennes.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : `https://github.com/Segolene-Albouy/Memoire-TNAH2019` (visité le 13/06/2022).

Chapitre 8

Visualisation et interfaces : un problème pour les humanités numériques ?

Ce chapitre de conclusion, plus théorique, revient sur les débats actuels concernant le lien entre interfaces graphiques, visualisation de données et humanités numériques. En s'appuyant (entre autres) sur les théories de Johanna Drucker [1] et Anthony Masure [2], il s'attache à réintégrer les questions « visuelles » propres aux humanités numériques à un contexte plus large. Le chapitre revient sur les origines des interfaces graphiques (et donc sur les objectifs implicites qui structurent nos interactions avec les machines) ; il cherche à remettre la visualisation dans les humanités numériques en lien avec une tendance globale à la visualisation – tendance qui vient du monde de l'entreprise, ce qui n'est pas anodin. Du fait de la quantité de significations implicites sous-jacentes à la construction d'interfaces, il est, je pense, nécessaire d'avoir une approche théorique et critique du rapport des humanités numériques aux problématiques de design et aux « dispositifs » [3] que sont nos outils de travail.

1. Johanna Drucker, *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

2. Anthony Masure, *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01).

3. Giorgio Agamben, *What is an apparatus ? and Other Essays*, Première édition, Stanford, 2009 (Meridian : Crossing aesthetics).

Bibliographie

À propos du projet Katabase / MSS

Corbières (Caroline), *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie) et Khemakhem (Mohamed), « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *IX Convegno AIUCD*, Milan, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02388407` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie), Gille Levenson (Matthias), Petkovic (Ljudmila) et Bartz (Alexandre), « Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? », dans *Digital Humanities DH2020*, Ottawa, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02898929/` (visité le 13/06/2022).

Janès (Juliette), *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

Khemakhem (Mohamed), Romary (Laurent), Gabay (Simon), Bohbot (Hervé), Frontini (Francesca) et Luxardo (Giancarlo), « Automatically Encoding Encyclopedic-like Resources in TEI », dans *The annual TEI Conference and Members Meeting*, Tokyo, 2018, URL : `https://hal.archives-ouvertes.fr/hal-01819505` (visité le 13/06/2022).

— « Information Extraction Workflow for Digitised Entry-based Documents. » Dans *DARIAH Annual event 2020*, Zagreb, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02508549` (visité le 13/06/1997).

Rondeau du Noyer (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Cha-*

ravay, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/lairaines/M2TNAH (visité le 13/06/2022).

Rondeau du Noyer (Lucie), Gabay (Simon), Khemakhem (Mohamed) et Romary (Laurent), « Scaling up Automatic Structuring of Manuscript Sales Catalogues », dans *TEI 2019 : What is text, really ? TEI and beyond*, Graz, 2019, URL : https://hal.archives-ouvertes.fr/hal-02272962 (visité le 13/06/2022).

Visualisation et design d'interfaces

Agamben (Giorgio), *What is an apparatus ? and Other Essays*, Première édition, Stanford, 2009 (Meridian : Crossing aesthetics).

Albouy (Ségolène), *Médiation des données de la recherche. Élaboration d'une plateforme en ligne pour une base de tables astronomiques anciennes.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/Segolene-Albouy/Memoire-TNAH2019 (visité le 13/06/2022).

Drucker (Johanna), *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

Masure (Anthony), *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01).

Économétrie et statistiques

Piketty (Thomas), *Les hauts revenus en France au XXe siècle : inégalités et redistributions, 1901-1998*, Paris, 2001.

Précision et rappel, Wikipedia. L'encyclopédie libre, 2022, URL : https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel (visité le 13/06/2022).

Table des matières